

# **Section 5**

## **Spatial Simulation**

# Smart Sampling

## Geostatistical Simulation

- *A probability-based technique (Monte Carlo process) on spatially correlated distributions*
- *Sacrifices the local best estimate for the reproduction of global statistics and features*
- *Simulation process can create any number of equally probable realizations (maps) all of which honor the available information*
- *Simulation allows for evaluation of joint uncertainty (accuracy) at multiple locations*

*Mound Accelerated Site Technology Deployment*

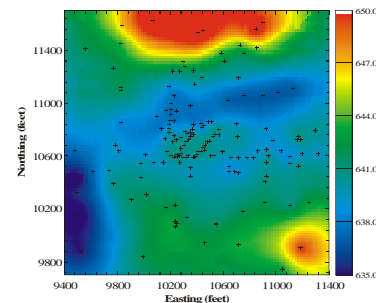
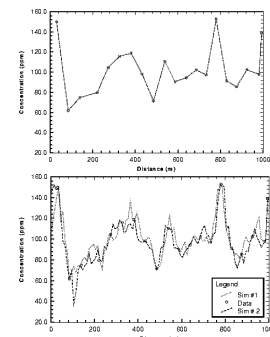
*5-2*

Recall the earlier example of walking across a site and collecting 20 samples along the transect. You could fit smooth lines through the points, which is like the kriging estimate ...

OR

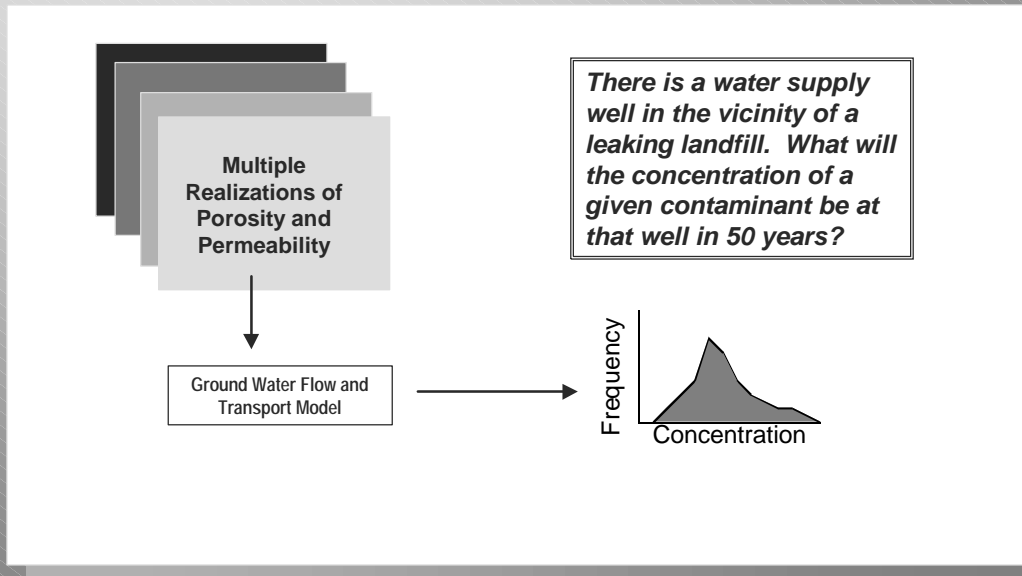
make multiple unique pictures from simulation, all of which honor the sample data.

- Accurately evaluating the best estimate at each location (kriging) produces an overly smoothed map. The kriged map displays a very smooth transition from red to blue to green with no sharp breaks. Many natural properties are not represented that way.
- By honoring the raw data values, histogram and variograms, simulation can retain the actual variability and better reproduce the global statistics and features.
- In kriging we're just trying to accurately evaluate each point. With simulation, we're trying to accurately model the global histogram and create images that reproduce larger scale features.
- The best map that we get out of kriging is probably not going to look like reality. Simulation provides a way to draw randomly from that best map to reproduce patterns or features that we believe exist in that data.



Kriged Map

- *Simulation provides a more realistic picture of natural complexity*
- *Simulation can provide an idea of “best,” “most likely” and “worst” cases for a given problem*
- *Simulation is a basis for Monte Carlo risk analysis where a full distribution of results is necessary*
- *Simulation reproduces the observed level of variability or heterogeneity at a site*



*Mound Accelerated Site Technology Deployment*

*5-4*

Simulation is used extensively in the petroleum, groundwater and nuclear waste industry. If your problem is soil only, simulation stops after generation of the contaminant distribution realizations. Design of the remediation scheme will proceed from an analysis of the variability and spatial distribution.

If your problem involves hydrology, simulation can be used to create multiple realizations of a property which can be input to a transfer function like a ground water flow and transport model, generating a probabilistic metric.

For the stated problem, porosity and permeability are spatially variable and uncertain. We have data from a few wells, and can build a variogram; we can generate maps, use them as input, and get a probabilistic metric for concentrations 50 years from now. Then it is up to the regulators and stakeholders to decide if the probability of high contaminants in the well is acceptably small. In the oil industry, where they want to know how much oil that can recover, the realization would be plugged into an oil reservoir model and produce a distribution of recoverable oil.

# Smart Sampling

## General Types of Simulation

**Parametric:** requires transform of the data to a parametric space, simulation in that space and then back-transform to raw data space.

**Example:** gaussian simulation using the normal-score transform

**Advantage:** only requires one-variogram model

**Disadvantage:** does not reproduce variogram at extremes of the distribution

**Non-Parametric:** requires discretization of data into classes and a variogram model at each threshold or class.

**Example:** indicator simulation of geologic facies (sand, silt, clay)

**Advantage:** Reproduces each variogram at each class/threshold

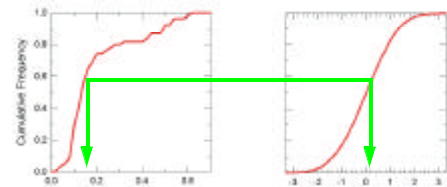
**Disadvantage:** requires variogram modeling for each class/threshold

*Mound Accelerated Site Technology Deployment*

5-5

**Parametric:** means there is some analytical expression that will describe the distribution. For a histogram, we want to say that the concentrations are normally distributed and may be fully defined by two parameters; the mean and variance.

Often, the plot of the discrete data doesn't follow a Gaussian distribution very well. The transform takes the raw distribution and maps it onto a Gaussian distribution with a mean of zero and a variance of 1 - a standard-normal distribution. From the data point at the 10<sup>th</sup> percentile of the actual distribution, we can draw a line to the 10<sup>th</sup> percentile of the Gaussian distribution where the mean and variance are known. This transforms the misshapen raw data distribution into a well-behaved, well understood distribution. After generating the simulation, we transform back to the raw data distribution; the mapping goes both ways.



The high and the low ends of the distribution won't show the spatial continuity that might actually exist. Though not critical for soil contamination problems, this can be a serious problem for groundwater problems.

**Non-parametric:** For non-parametric simulation, establish threshold points, designating a value of 1 if the point is less than or equal to the threshold and 0 if otherwise. Determine the percentage of data that meets each threshold, then perform the transformation and create the variogram for each threshold.

# Smart Sampling

## Simulation Algorithms

*Turing Bands:* late 1960's, extra work to condition to data

*LU Decomposition:* good for small domains with few conditioning data

*Sequential:* currently popular, conditioning to data by construction

*Probability Field:* apply spatially correlated random numbers to set cdf field

*Simulated Annealing:* perturb field of values until they match a defined metric.

*Mound Accelerated Site Technology Deployment*

5-6

In geostatistics we typically use the sequential simulation algorithm because it conditions every simulation to the available data as it goes, and never lets the data values change.

# Smart Sampling

## Sequential Simulation

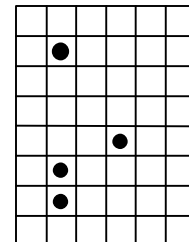
- *Map the conditioning data onto a grid*
- *Randomly visit all other grid nodes*
- *Use kriging system to create a local cdf based on surrounding data for each node*
- *Draw a random value from the cdf to get the simulated value at that location*
- *Consider each simulated point as a conditioning value for future cdf construction*
- *Continue until all nodes have a simulated value*
- *Reinitialize random number generator and begin next realization*

*Mound Accelerated Site Technology Deployment*

*5-7*

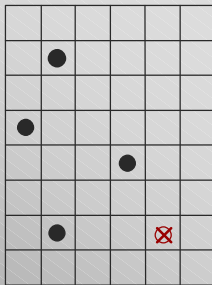
● indicates an actual sample value.

We want to fill in the entire grid through sequential simulation.



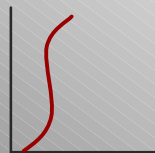
# Smart Sampling

## Sequential Simulation Example

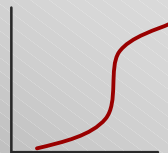


- Use the kriging system to create a local cdf based on the surrounding data points of the first node.
- Develop a cdf for this location.
- Draw a random number between 0 and 1, and assign the value for that probability to the node.
- For the remainder of this realization, the newly defined node is treated as a sample point.

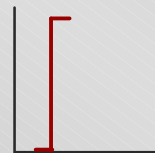
*Local cdfs reflect proximity to data locations*



*Well defined  
Close to data*



*Poorly defined  
Far from data*

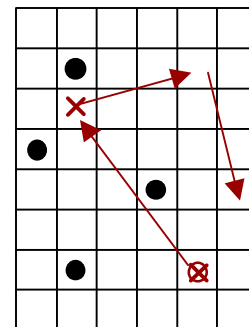


*On top of / right next to value  
On top of data point*

*Mound Accelerated Site Technology Deployment*

5-8

- Randomly jump to another node and repeat the process, this time including the node just calculated as one of the surrounding data points.
- Keep jumping around randomly until the whole grid is filled (never visit a node more than once).
- To make the next realization, use a different seed and follow a different path to fill the grid.



The kriging system gives us the best estimate, which is the mean at that location if we're using parametric simulation; and the variance is from the kriging variance.

At that first point, the distribution is fairly broad because we're far away from any data point. At the second point it will be a tight distribution, heavily weighted by the two points close to it. Without a broad range of possible values, if we draw .3 or .7 probability, the value at the second unknown point will be fairly close to the values of the close data points.

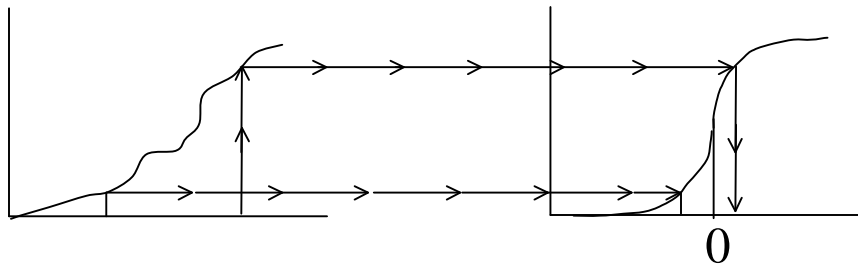
**Gaussian:** normal-score transform allows the kriging estimate and kriging variance to define the local cdf.

**Indicator:** construct the cdf through indicator kriging at each threshold  $z_k$ .

The expected value of the cdf at any threshold is estimated by the weighted linear combination of surrounding indicator data.

Smart Sampling uses the Gaussian parametric because it allows modeling of many different action levels without having to redo the variogram.

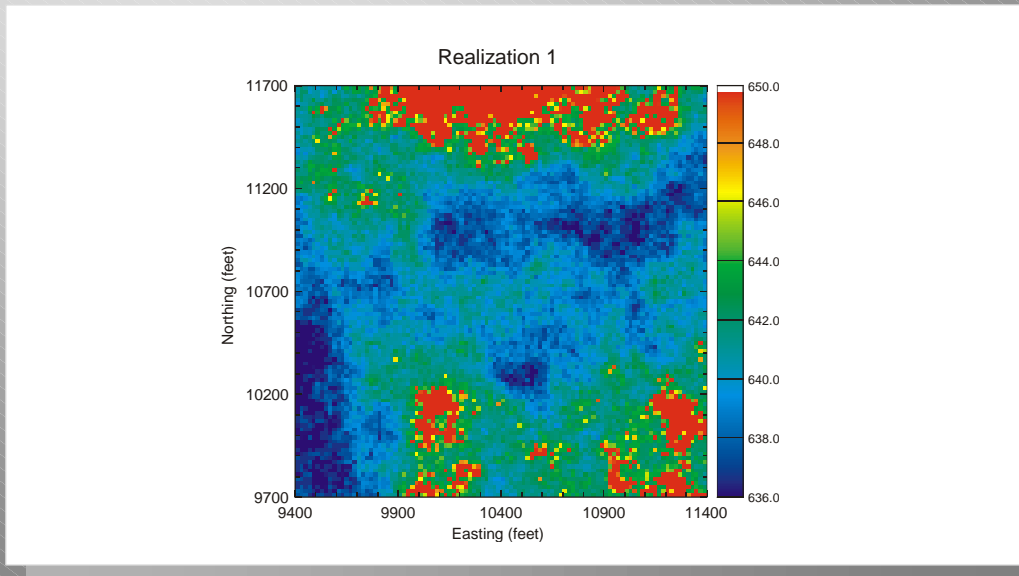
Transform the raw data into a Gaussian distribution with mean = 0 and variance = 1. The Gaussian model provides a weighted average of the surrounding points



For Indicator Simulation modeling, we construct the cdf as we go for each data point. Instead of coming up with a different Gaussian distribution at each point and drawing randomly from that, we're constructing the distribution at every point.

$$\frac{1}{n} \sum w_i z_k(i)$$

This way takes longer, because we have to solve the kriging equation 5 times before doing the random draw.



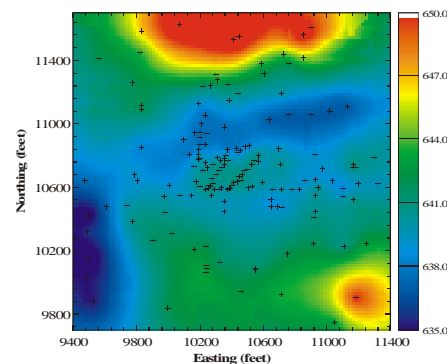
*Mound Accelerated Site Technology Deployment*

*5-10*

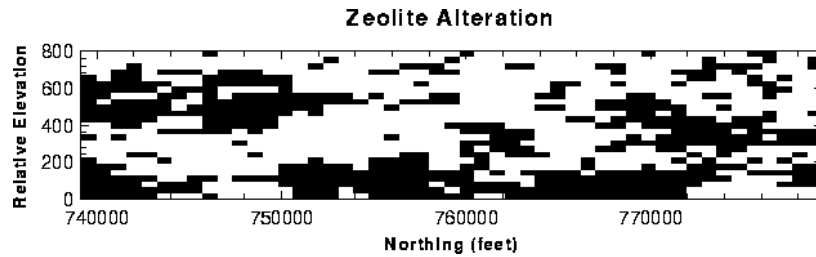
When compared to kriging map (right), the realization (above) displays much more variability.

Another way to look at the relationship:

If you took an infinite number of realizations (maps like this) and stacked them all up on each other, took the average value across the stack of maps at each pixel, and mapped that average value, you'd get the kriged map.



**SEAN:** Suggestion was to show histogram of data vs. histogram of simulation, perhaps both cdfs as well. If we no longer have this realization (HA!), maybe we could use something you've already done on Rocky or other site? What site/data is this kriged map from? (original is from GSA section 4 slide 23)



*Mound Accelerated Site Technology Deployment*

*5-11*

The threshold was applied to the data and then simulated the values that were above and below the data.

In each realization the data plotted is binary -- 0 or 1 relative to threshold.

Black is below threshold, white above.

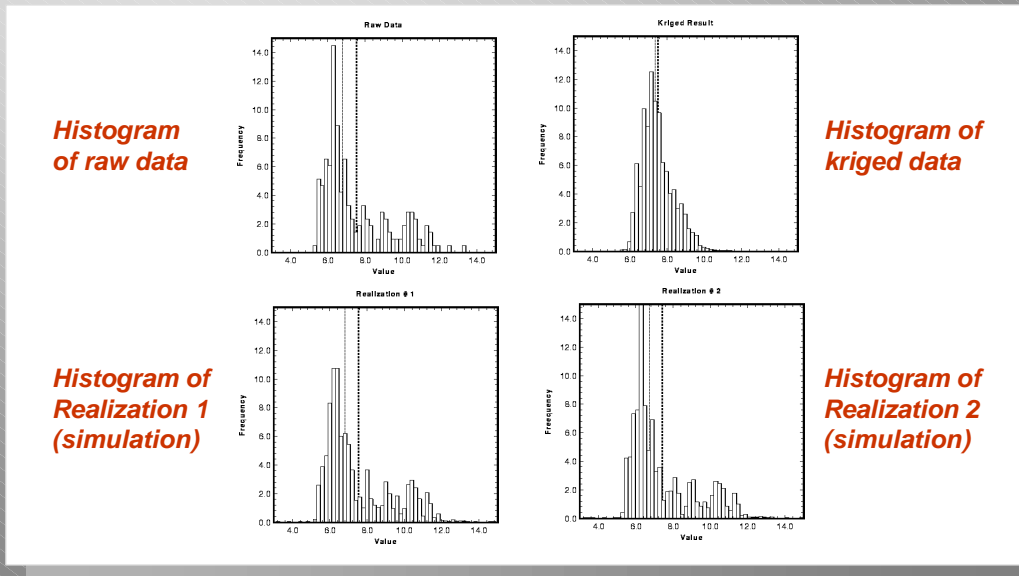
**Kriging:** *smoothing effect of interpolation will produce:*

- 1) *A longer range variogram in the output than the input model*
- 2) *Less variability in the output field than the input data (distribution gets squeezed)*

**Simulation:** *attempts to reproduce the input histogram and variogram (the input univariate and bivariate data distributions, respectively) within the limits of “ergodic fluctuations”.*

The variogram of the kriged data shows more correlation than the variogram of the raw data. This increased range is a product of kriging's smoothing effect. The kriged data will have a tighter cdf.

Every single one of the realizations will not point-for-point match the raw variogram exactly, but on average across all of the realizations, they will match the distributions.



*Mound Accelerated Site Technology Deployment*

*5-13*

These graphs show that the mean was reproduced both by kriging and by simulation. The median is better reproduced by the realizations.

The distribution of kriged values is much smoother, centering the data on the mean. The input values are represented but none of the many additional points are at the extremes.

The histograms of values from two realizations created by simulation do a much better job at reproducing the raw data. A realization will not reproduce the raw data exactly, but an average of 100 or so realizations should be very close to the statistics of the raw data.

- For the standard deviation, the spread of the distribution, kriging only shows about half of the variability of the raw data while the simulations reproduce it pretty well.
- The realizations show possible highs and lows that were not captured in the sampling.

Kriging reduces variance but retains the mean of the input data. As an interpolator, it does not produce values outside the minimum and maximum of the sample data.

Parameter	Raw Data (n=214)	Kriged Map (n=5329)	Realization 1 (n=5329)	Realization 2 (n=5329)
Mean	7.53	7.49	7.41	7.52
Median	6.79	7.34	6.70	6.81
Standard Deviation	1.81	0.86	1.75	1.80
Minimum	5.20	5.46	3.18	3.10
Maximum	13.24	13.13	14.93	14.88
10th Percentile	5.80	6.46	5.80	5.80
90th Percentile	10.47	8.75	10.39	10.47

*Mound Accelerated Site Technology Deployment*

*5-14*

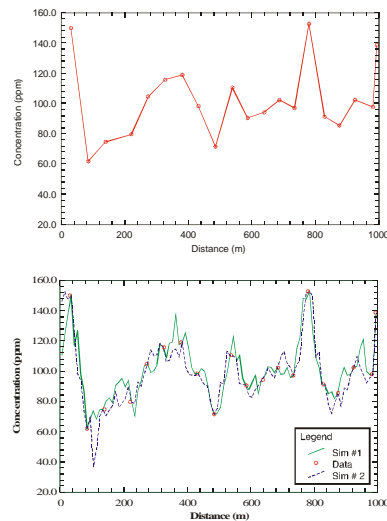
In this example, 214 raw data points were used to create a kriged map and two realizations. The mean was reproduced in all of them; the median is better reproduced by the realizations.

The standard deviation (the spread of the distribution) looks better with simulation; the kriged data showing only about half of the variability of the raw data. Ergodic fluctuations can be seen.

# Smart Sampling

## Notes on Univariate Distributions

- **Kriging** reduces variance but retains the mean of the input data
- **Kriging**, as an interpolator, does not produce values outside minimum and maximum of sample data
- **Simulation** can produce values above and below the maximum and minimum sample data because it draws from a fully defined cdf  $[0, 1]$  at each location.



*Mound Accelerated Site Technology Deployment*

*5-15*

Talk about relationship of two techniques to the minimum and maximum samples.

- *Generally, both the absolute minimum and maximum aren't randomly included in a limited data set*
- *It is possible to simulate values outside the sample range by defining the data distribution beyond the sample value extremes*
- ***add graphs / charts from GSLIB Ver 2 p.134-138***
- *Use a power function or hyperbolic model to do this extrapolation in GSLIB software*
- *Extrapolation can also be truncated by setting a minimum and maximum possible value.*

You can use the variability within the sample data to help determine how far outside the measured maximum and minimum to extend the cdf.

Also, use whatever knowledge of the distribution or the property you have. For example, for porosity the values must be between 0 and 1.

Air pollution modeling is a classic example for truncating the extrapolation; you want to define particulate size - restrict it so that there are no negatives and to prevent extending the tail all the way to include boulders.

- *Ergodic fluctuation is defined as the difference between the input model and the statistics of a realization.*
- *Input models are generally based on data from a limited sample size*
- *The underlying model is said to be ergodic in the parameter  $a$  if the realization statistics tend toward  $a$  as the size of the field increases*

*Mound Accelerated Site Technology Deployment*

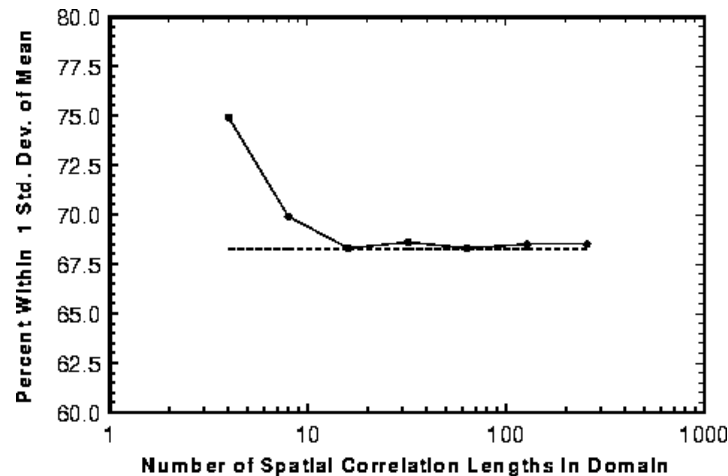
*5-17*

As pointed out with the table on comparing univariate statistics, each simulation is not going to exactly reproduce the mean, median, and variance, but they will be close.

We don't want to reproduce the input data exactly because these models are based on data from a limited sample size. Just because we took 100 samples, why should we require all realizations to tie into those parameters? If we had taken 200 samples, those parameters may have been different.

Parameter	Raw Data (n=214)	Realization 1 (n=5329)	Realization 2 (n=5329)
Mean	7.53	7.41	7.52
Median	6.79	6.70	6.81
Standard Deviation	1.81	1.75	1.80
Minimum	5.20	3.18	3.10
Maximum	13.24	14.93	14.88
10th Percentile	5.80	5.80	5.80
90th Percentile	10.47	10.39	10.47

If it were possible to sample on an infinite field, taking more and more samples, the sample statistics would tend towards the actual population's statistics.



*Mound Accelerated Site Technology Deployment*

*5-18*

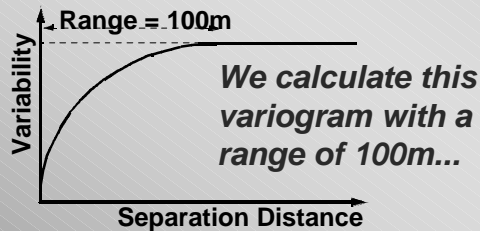
The practical manifestation of ergodic fluctuation can be shown through the following process.

- Create a number of unconditional realizations. A Gaussian distribution will have exactly 68.3% of the values between  $\pm 1$  standard deviation of the mean.
- Create models which have different variogram lengths relative to the domain size; basically the number of ranges that fit within the domain.
- A very small range, relative to the domain size, might fit 100 or 300 variograms and will reproduce the theoretical statistics fairly well. That holds true until the domain size becomes less than 10 correlation lengths, then the sample statistics are no longer represented.
- Ergodic assumption: Domain size must be  $\geq 8-10$  correlation lengths
- In practice, we often have variogram ranges close to the domain size or even half of the domain size. This is mitigated considerably by having conditioning data (samples) which can constrain the simulations across the site.
- Domain size usually derives from the site, a specific size area to look at.
- Correlation lengths come from the sample data, and depend on the deposition mode.

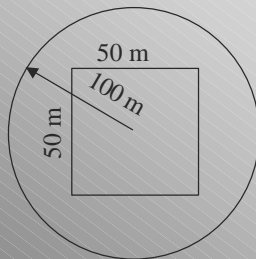
# Smart Sampling

## Ergodic Assumption

*Domain size must be <sup>3</sup> 8-10 correlation lengths*



*... but we have a simulation domain that is only 50m x 50m.*



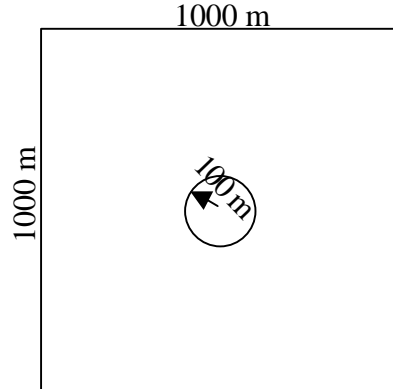
*If there are no conditioning data, any simulated point will be heavily weighted by the first point because the variogram range extends over the entire domain.*

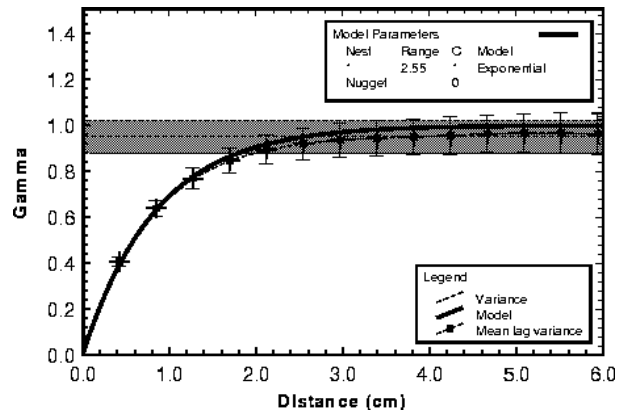
*Mound Accelerated Site Technology Deployment*

*5-19*

If the first point is a high point, the whole area will get filled in with high values, it's not going to reproduce the sample statistics. If a very low value is simulated first, the whole area will be low values.

If we make the domain size 1000x1000, the variogram fits in the domain size fairly easily and will reproduce the statistics.





*Mound Accelerated Site Technology Deployment*

5-20

You can check your output by comparing bivariate statistics. After modeling a variogram from the sample data (solid black line) and creating 100 realizations, recalculate the variogram in the same direction with the same search parameters for each one of those realizations.

- The gray band shows the total variance across the 100 realizations.
- The error bars show the 95% confidence levels for every variogram lag.
- The black squares show the mean of the hundred variograms.

The realizations captured the actual input variogram at pretty much every lag. Although still within the (?), the realizations underpredict the mean.

Chris would like to see histograms for this slide. Sean would too, but does not have software to do this easily.